Université de Caen Normandie

IUT Grand Ouest Normandie - Pôle de Caen

Département Science des Données

Diplôme Bachelor Universitaire de Technologie

SCIENCE DES DONNEES

PORTFOLIO

Lou-Anne Thomas

Année universitaire 2022-2025

Table des matières

SAÉ 1-01 Création de reporting à partir de données stockées dans un SGBD relationnel

A. Présentation du projet

Lors de ce projet, nous avions à notre disposition un jeu de données fictives concernant une ville prénommée « Gotham ». Dans un premier temps, nous avions pour mission de **retrouver l'identité** de celui qui se fait appeler « **Batman** » parmi plus de cinquante mille individus en s'aidant uniquement du langage **SQL**. Pour cela, différents **fichiers au format CSV** ainsi qu'une **base de données** (cf. Figure 1) étaient à notre disposition. En second lieu, nous avons travaillé sur **la représentation graphique** des données de la ville de Gotham. Un rapport retraçant ces deux parties était requis.

B. Organisation

Afin de réaliser ce travail, des groupes ont été composés. Nous étions ainsi trois à travailler sur ce projet. La première étape fut l'étude des données ainsi que la mise en place de règles à suivre. Nous avons choisi d'effectuer la SAE en **SQLite** pour des raisons pratiques. En effet, SQLite est plus portable que MySQL, bien que ce dernier soit plus puissant.

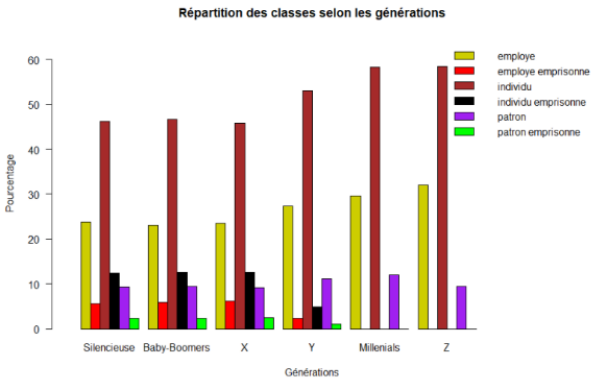
C. Démarche et résolution du problème

1. Conception de nos bases de données : Tout d'abord, nous avons dû **lire les fichiers CSV** fournis en utilisant Python. Par la suite, nous avons **créé une base de données** supplémentaire à celle déjà fournie dans laquelle nous avons **créé des tables** et **inséré les données récupérées** à partir des fichiers CSV. Pour accomplir cette tâche, nous avons écrit des requêtes SQL et manipulé différents types de données tels que des chaînes de caractères, des nombres et des dates.

2. Enquête : Nous disposions d'indices pour identifier Batman parmi ces milliers d'individus. En les combinant, nous avons traduit ces indices en quatre requêtes SQL distinctes. Ces requêtes ont nécessité la manipulation de différentes fonctions SQL et la réalisation de **jointures entre les tables**. Nous avons ensuite récupéré les résultats dans des variables python sous forme de liste, puis comparé les identifiants pour n'en garder qu'un seul, celui de Batman.

3. Représentation graphique : Dans cette partie, le maire de Gotham nous a demandé de travailler sur la représentation graphique des données. Pour réaliser le graphique (cf. Figure 1), nous avons dû créer une nouvelle catégorie appelée "génération". Pour ce faire, nous avons écrit une liste python pour chaque génération, puis nous avons classé les individus en fonction de leur année de naissance et les avons attribués à la bonne liste. Enfin, nous avons compilé toutes les données dans un fichier csv pour réaliser les graphiques en utilisant la **librairie pandas** en Python.

Figure 1 – Graphique réalisé à l'aide de la librairie Pandas



D. Conclusion

Ainsi, de nombreux bénéfices ressortent de ce projet. En effet, c'était la première fois que je travaillais sur des données aussi nombreuses. Cela m'a fait comprendre la nécessité d'être rigoureuse dans l'écriture de requêtes et d'optimiser ces dernières afin de minimiser l'attente liée au requêtage. J'ai également appris auprès de mes camarades à respecter les usages en matière d'architecture de dossier et à organiser de façon claire mes fichiers. D'autre part, ce projet fut une introduction à la représentation de données et nous a permis d'exploiter nos acquis sur différents logiciels et langages : Excel, Python, SQLite.

SAÉ 1-02 Écriture et lecture de fichiers de données

A. Présentation du projet

Le collège Ribery, collège fictif, est une école à **2 niveaux** (Licence 1 et Licence 2) où un niveau se décompose en **quatre classes**. C'est suite à la demande de la directrice du collège que ce projet a lieu. En effet, les notes des étudiants sont dispersées dans plusieurs fichiers (cf. Figure 2), nous devions ainsi être en capacité de **rassembler l'ensemble des notes au sein d'un seul document**, de **déterminer** quels sont les **points fort** et **points faibles** de chaque promo, mais aussi d'**éditer un carnet de notes** (au format **txt**) pour chaque élève.

Figure 2 - Extrait de trois des fichiers à notre disposition

B. Organisation

Pour réaliser ce projet, des groupes ont été composés. Nous étions ainsi deux à exploiter ces données à l'aide du langage de programmation **python**. La première étape fut **l'étude des données**, elle nous a permis de nous faire une vision globale du projet et de lister les traitements dont nous avions besoin de réaliser. Il a également fallu se coordonner et mettre en place des règles à suivre, telle que la mise en place d'une nomenclature pour le nom de nos variables python. La seconde étape fut la **création d'une liste python**. En effet, nous avons décidé de recenser l'ensemble des données sous le format d'une liste python qui nous permettra **d'écrire un fichier CSV** que l'on a pu facilement **manipuler sur Excel**.

C. Démarche et résolution du problème

Afin d'établir cette liste python, nous avons suivi le protocole suivant :

- Lecture et récupération des données du fichier dans une liste python.
- Traitement des caractères si nécessaire.
- Trie par ordre alphabétique des noms de la liste.

La structure des fichiers n'étant pas homogène (cf. Figure 2) , il a été nécessaire de les traiter au cas par cas. Ce projet nous a ainsi permis de **manipuler des chaînes de caractères en python** : ajout d'éléments séparateur, suppression des espaces, des lignes vides et des caractères non désirés.
L'obtention de notre liste nous a permis d'éditer des carnets de notes pour chaque élève (cf. Figure 3). Nous avons par la suite, à l'aide de python, écrit un fichier CSV (cf. Figure 4) que nous avons importé sur Excel pour faire l'analyse des points forts et points faibles de chaque promo.

Figure 3 - Extrait d'un carnet de notes

```
AGUILAR Felipe

Moyenne générale : 13.833
Mention : admis

Maths    Français SVT      Phy-chi  Anglais  Sport
20       12       7        12       19       13
```

Figure 4 - Extrait du fichier csv réalisé

```
licence,promo,nom,prenom,SVT,Physique-Chimie,Sport,Maths,Anglais,Francais
L1,L1P2,Aguilar,Felipe,7,12,13,12,19,20
L2,L2P3,Alexander,Julie,0,8,11,2,17,14
L2,L2P1,Alteri,Shawn,18,6,14,11,7,20
L1,L1P2,Bailey,Daniel,0,6,11,18,20,13
L2,L2P2,Bickel,Marty,11,15,5,11,20,19
L2,L2P1,Blackman,Adolph,17,13,18,20,13,20
```

D. Conclusion

Ce projet fut une première approche à un langage informatique. Il nous a permis d'apprendre à traiter des chaînes de caractères, mais aussi à lire et écrire des fichiers au format TXT et CSV. Le choix d'effectuer le projet sous forme de liste python nous a familiarisé à l'utilisation de boucles python et des index de listes python. C'était par ailleurs le premier projet contenant du code, j'ai ainsi découvert l'existence des packages python et compris l'importance d'avoir un code clair et commenté afin de gagner du temps dans la recherche d'erreur.

SAÉ 1-03 Préparation et synthèses d'un tableau de données en vue d'une analyse exploratoire simple

A. Présentation du projet

Dans le cadre de cette SAE, nous avons réalisé à l'aide du **langage de programmation statistique R**, une analyse exploratoire portant sur l'utilisation des vélos dans la ville de New York au cours du mois de juin 2022. Nous avions ainsi à notre disposition un jeu de données regroupant les locations de vélos du groupe City Bike. Ce fichier comportait plus de trois millions de données que nous avons traités pour mieux comprendre les habitudes et les comportements des cyclistes.

B. Organisation

Le projet s'est déroulé en deux phases, la première consistait à **explorer les données**, c'est-à-dire les comprendre, les nettoyer et créer des visuels en R. La seconde phase, quant à elle, consistait à **résumer** et à **interpréter** nos sorties créées en R au sein d'un rapport. Pour réaliser ce projet, nous étions deux à travailler sur les données.

C. Démarche et résolution du problème

Dans une première partie, nous avons préparé les données en ciblant les variables pertinentes à notre analyse et en effectuant les traitements nécessaires. Nous avons par exemple, ajouté une colonne durée de location à partir des variables de début et de fin de location que nous avions au préalable transformés dans le bon format de date. Lors de la seconde partie, nous avons effectué le reste du code portant sur l'analyse des données. Ainsi, nous cherchions sous différents angles à résumé nos variables pour mieux les comprendre et pour en déduire les données aberrantes ou atypiques potentiellement présentes. Ces valeurs pouvaient être, des données négatives dans la variable de durée de location (valeurs aberrantes) ou encore des valeurs extrêmes tels que des durées supérieures à 60 min (atypiques).

Figure 5 - Histogramme de la variable de duration avec les valeurs atypiques

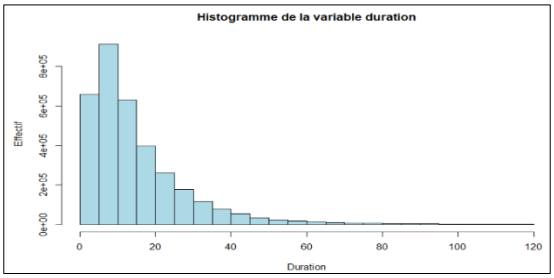
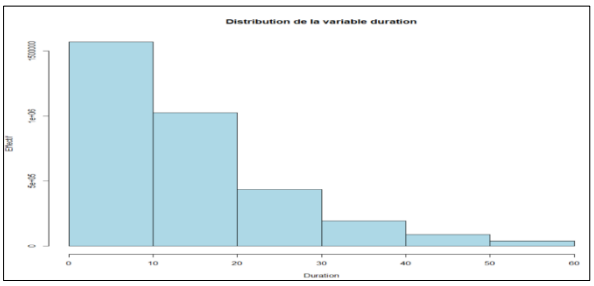


Figure 6 - Distribution de la variable de duration sans les valeurs atypiques



Concernant la variable « duration », il était facile d'identifier les valeurs aberrantes, ces dernières étaient négatives, cependant, il était plus difficile d'identifier les valeurs atypiques, nous avons donc dressé l'histogramme présent sur la figure gauche (cf. Figure 5). Ainsi, on remarque que les valeurs supérieures à 60 minutes sont négligeables, nous les avons donc éliminées. En se référent à la figure de droite ci-dessus (cf. Figure 6), on remarque que les données sont plus claires et plus détaillées qu'auparavant.

D. Conclusion

En conclusion, nous avons pu étudier sous différents axes le comportement des utilisateurs. Nous en avons conclu que les utilisateurs préféraient les vélos classiques et effectués des trajets de courte durée, que le mardi et vendredi comportaient le plus grand nombre de locations et que les stations les plus populaires étaient celles situées à proximité de grands parcs et aéroports. Cette première SAE en R, nous a également permis de découvrir ce langage et de nous y familiariser. Nous avons aussi pu mettre en pratique les notions vues lors du module de statistiques descriptives, telles que l'identification du type des variables et la présentation des données sous le graphique approprié pour fournir une analyse personnelle.

A. Présentation du projet

Pour ce projet, nous devions créer un **poster scientifique** résumant le **traitement de l'information au sein d'une organisation**. Pour cela, nous avons mené des **recherches documentaires** au sein de presse économique, de gestion et de marketing. Ce travail nous a permis de comprendre les notions relatives aux technologies de l'information et à l'informatique décisionnelle tout en exigeant une **capacité de synthèse** pour présenter un résumé complet de nos recherches.

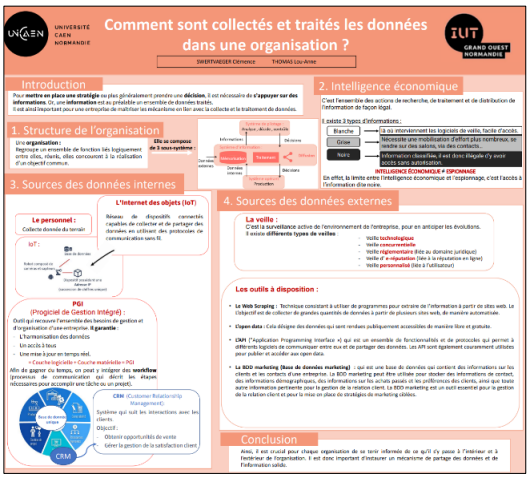
B. Organisation

Nous avons travaillé en binôme pour ce projet qui s'est déroulé en deux parties. La première était dédiée à la construction de l'information du poster. Cela consistait à rassembler et à trier les informations nécessaires pour le poster. La seconde partie, quant à elle, s'est dédiée à la mise en forme du poster. Nous avons donc établi des critères importants tels que l'utilisation de phrases courtes, l'insertion de schémas, la création d'un contraste entre le fond et la police d'écriture, ainsi que l'utilisation d'un code couleur cohérent.

C. Démarche et résolution du problème

Comme mentionné précédemment, la première étape fut la prise d'information au sein de diverses sources telles que des articles de presse, de site internet ou encore de vidéos. En rassemblant l'ensemble de nos recherches, nous avons listé au brouillon les notions à aborder. Le contenu principal du poster se divise donc en deux catégories : les **sources internes de collectes** de données et **les sources externes**.

Figure 7 - Poster réalisé



La seconde étape fut la création des différentes parties autour d'une même ligne directrice, nous avons ainsi décidé d'expliquer les notions générales au début telles que la structure d'une organisation et la notion d'intelligence économique pour entrer, par la suite, dans les détails techniques de la collecte des données. Ainsi, le poster suit une suite logique, en mettant les notions les plus importantes au début.

Les parties de notre brouillon structurées, nous avons travaillé sur la mise en forme et la disposition de nos parties. Pour réaliser le poster, nous nous sommes ainsi inspiré de différents posters scientifiques trouvés sur Internet.

D. Conclusion

Ce travail m'a permis de mieux comprendre les enjeux et les outils liés à l'**intelligence économique** et à **la veille**. Ces apprentissages faciliteront également la recherche de futurs stages, en effet, être familiarisé avec ces notions démontre aux futurs recruteurs un intérêt pour ce domaine. À travers cette SAE, j'ai compris l'importance de mettre en place un mécanisme de partage de l'information et de traitement des données au sein d'une organisation. Il est également important pour une organisation de se tenir informée de son environnement interne et externe, cela lui permet de ne pas se déconnecter de son environnement et, ainsi, de prendre des décisions éclairées.

A. Présentation du projet

Pour ce projet, nous avons travaillé sur la **réalisation d'un poster** présentant **la place des femmes dans le milieu professionnel en Normandie**. Une **présentation orale** a par la suite eu lieu **en anglais**. Ce travail m'a permis de me familiariser avec des termes anglais liés aux statistiques, mais aussi de m'informer sur les filières d'excellences normandes, telles que l'agro-alimentaire et l'énergie.

B. Organisation

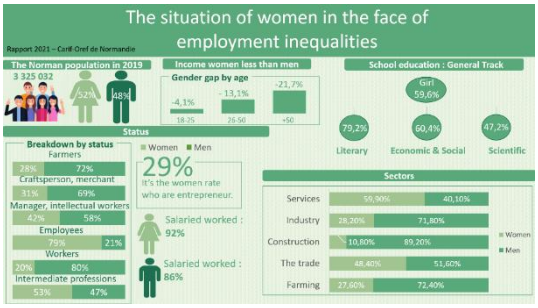
Nous avons travaillé en binôme sur ce poster. Sa création a nécessité deux phases. Une phase documentaire afin de nous informer sur la situation des femmes dans le milieu professionnel en Normandie, et une phase de conception graphique pour mettre en forme les informations collectées. Nous avons effectué des recherches documentaires chacune de notre côté, puis nous avons procédé à une séance de brainstorming pour mieux définir les points clef à aborder dans notre travail. Cette démarche nous a permis d'avoir une vision plus claire et plus approfondi de notre sujet.

C. Démarche et résolution du problème

1. Conception du poster :

Nous avons choisi d'utiliser un rapport publié par Carif-Oref de Normandie en 2021 pour la réalisation de notre poster. Ce rapport couvre plusieurs thématiques telles que la population, l'emploi, le marché du travail, la formation et l'entrepreneuriat.

Figure 8 - Poster réalisé



Nous avons examiné ces sujets en utilisant différents indicateurs que nous avons identifiés lors de la phase de brainstorming en définissant quatre axes d'analyse : la différence de statut entre les hommes et les femmes, la différence de salaire, les secteurs d'activité où les hommes et les femmes sont présents et les différences en termes d'éducation. Nous avons également effectué des recherches de vocabulaire pour comparer les termes relatifs aux statuts en France et dans les pays anglophones, ainsi que pour enrichir notre vocabulaire dans les domaines des statistiques et du monde professionnel en général.

En dernier lieu, nous avons travaillé sur la conception du poster, où nous avons établi un code couleur, créé des graphiques sur Excel et disposé les informations de manière à mettre en évidence les données les plus importantes.

2. Présentation orale :

Le projet s'est terminé sur une présentation orale en anglais. Nous avons présenté les différents indicateurs du poster. Une séance de question a ensuite eu lieu afin de nous faire échanger en anglais.

D. Conclusion

Ainsi, ce travail nous a permis d'enrichir notre vocabulaire dans le domaine des statistiques et du monde professionnel. Il a également mis en exercice notre esprit de synthèse et nous a permis de mieux comprendre la situation des femmes dans le monde professionnel en Normandie, ainsi que les obstacles qu'elles rencontrent.

A. Présentation du projet

Lors de ce projet, nous avons travaillé sur la **mise en œuvre des différentes étapes d'une enquête**. Nous avons **rédigé un questionnaire** et **effectué une analyse statistique** sur des données tirées d'une enquête de satisfaction réalisée par le Crédit Agricole Normandie.

B. Organisation

Nous avons travaillé en groupe de quatre pour élaborer un questionnaire portant sur les habitudes alimentaires des lycéens. Par la suite, nous avons travaillé sur les données de l'étude réalisée par le Crédit Agricole Normandie. Cette analyse s'est effectuée en utilisant le logiciel RStudio. L'ensemble de nos résultats ont été présentés dans un rapport et lors d'un entretien.

C. Démarche et résolution du problème

La rédaction du questionnaire (cf. Figure 9) nous a permis de nous confronter aux défis que demande l'élaboration d'un questionnaire. Cela inclut la nécessité de penser à la structure du questionnaire, de formuler des questions claires et pertinentes, de minimiser les biais de réponse et de trouver un équilibre entre la collecte de données nécessaires et la fatigue que peuvent éprouver les répondants. En outre, cela nous a amenés à élaborer un dictionnaire de variables pour coder les réponses du questionnaire. Pour élaborer notre questionnaire, nous avons dû réfléchir à un thème qui soit adapté à notre cible et structurer les sous-parties du questionnaire en conséquence. Ce processus n'était pas évident, car il était important de s'assurer que les résultats ne soient pas biaisés. Nous avons dû réfléchir à l'ordre des questions et les rendre accessibles à tous les répondants ciblés.

Figure 9 - Extrait du questionnaire

Les lycéens et l'alimentation

PARTIE A – Informations générales

Q.1 – Etes-vous un/une : ☐ Homme ☐ Femme

Q.2 – En quelle classe ? ☐ Seconde ☐ Première ☐ Terminale

Q.3 – Quelle filière envisagez-vous après le bac ? (Une seule réponse)

☐ Arts – Lettres – Langues ☐ Droit – Economie – Gestion

☐ Sciences Humaines et Sociales ☐ Sciences – Technologies – Santé

☐ Autres : .....

PARTIE B – Les habitudes

Q.4 – A quelle fréquence consommez-vous des plats préparés ?

Le Crédit Agricole Normandie a profité d'un entretien entre un conseiller et un client pour solliciter ses clients dans le cadre d'une enquête de satisfaction. Cette sollicitation a donné lieu à la création d'un fichier de données contenant les réponses aux différentes questions. L'objectif a alors été d'analyser ces données en vue d'apporter une réponse à différentes questions d'intérêt. Cette analyse s'est faite sur le logiciel RStudio. Nous avons créé différents tris à plat et graphiques que nous avons présentés dans un rapport et lors d'une soutenance. Ainsi, nous avons pris connaissance du questionnaire de l'étude, nettoyé les données et analysé les réponses sous différents angles.

D. Conclusion

Ce projet nous a permis de comprendre les différentes étapes de la création d'une enquête, de la formulation des questions jusqu'à la présentation des résultats. Nous avons également appris à utiliser le logiciel RStudio pour l'analyse des données. Nous avons rencontré différentes problématiques durant ce projet liées à la création d'un questionnaire adapté et à la manipulation des données. Ce travail a donc été l'occasion de mobiliser nos compétences en matière de rédaction, de traitement des données et de présentation.

A. Présentation du projet

Sur ce projet, nous avons travaillé sur **une analyse de régression linéaire simple** qui cherche à déterminer le meilleur modèle pour prédire l'indice de masse maigre. L'échantillon contenait 252 êtres humains de sexe masculin. Ainsi, nous avons étudié différentes variables morphologiques telles que la taille, le poids ou le tour de taille dans l'objectif de prédire la variable de masse maigre.

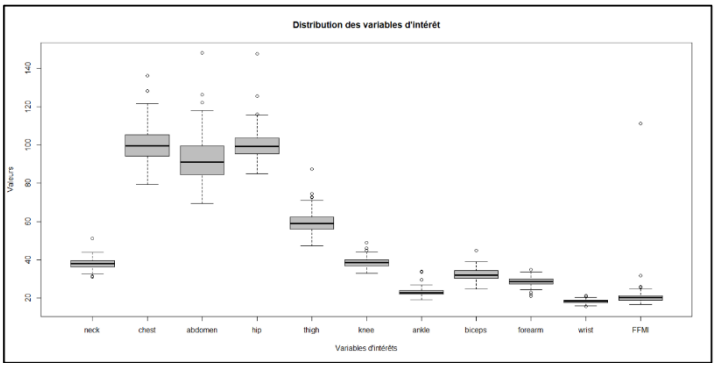
B. Organisation

En groupe de trois, nous avons réalisé cette analyse sur **Rstudio** à l'aide des **librairie flextable** pour représenter des tableaux et la **librairie modelsummary** pour visualiser les résultats des modèles. Ainsi, nous avons dans une première partie préparé et nettoyé nos données pour dresser par la suite, des modèles de régression linéaire croisant la variable de masse maigre à chaque variable morphologique. Dans une seconde partie, nous avons rassemblé l'ensemble de nos analyses dans un rapport et présenté notre démarche et nos résultats lors d'une soutenance.

C. Démarche et résolution du problème

Tout d'abord, nous avons travaillé sur la préparation des données. Pour cela, nous avons converti les données dans les unités nécessaires à la réalisation de nos calculs d'indicateurs de la variable de masse maigre. Nous avons recherché la présence de valeurs atypiques et aberrantes. Pour cela, nous avons visualisé la répartition de nos variables d'intérêts (cf. Figure 10) afin d'identifier les valeurs atypiques puis visualisé leur influence via des nuages de dispersion croisant les variables morphologiques à la variable de masse grasse.

Figure 10 - Graphique représentant la répartition des variables d'intérêts de notre projet



Après élimination de nos valeurs fortement atypiques qui influençaient nos résultats, nous avons **mesuré les associations linéaires** entre les variables morphologiques et la variable de masse grasse. Nous avons terminé par **dresser** nos **modèles de régression linéaire** et conclu sur le modèle le plus fiable en **mesurant leurs coefficients de détermination**. Ainsi, nous avons pu interpréter les coefficients de régression linéaire pour comprendre comment la variable la plus prédictrice influence la variable masse maigre.

D. Conclusion

En conclusion, cette analyse de la masse maigre nous a permis de mettre en pratique les notions de la régression linéaire que nous avons étudiées. Nous avons également pu constater l'importance de la préparation des données et de la répercussion d'une mauvaise gestion des données sur les résultats de nos modèles.

SAÉ 3.01 - Recueil et analyse de données par échantillonnage ou plan d'expérience

A. Présentation

J'ai effectué ce travail seul sur un créneau horaire défini, totalisant 4 heures, au cours desquelles j'ai mis en place un plan d'expérience à l'aide du logiciel **RStudio**. L'objectif de ce projet était de déterminer la composition optimale des ingrédients d'une pâte à pizza afin de maximiser sa qualité gustative. Pour ce faire, nous disposions de scores attribués à chaque recette variant selon les quantités de farine, de sel et de levure. L'analyse visait donc à déterminer les proportions idéales de chaque ingrédient influençant le goût de la pizza.

B. Démarche et résolution du problème

La démarche de l'étude a impliqué la définition de niveaux haut et bas pour les facteurs clefs de la pizza (farine, sel, levure). Pour chaque recette, deux répliques de la pâte ont été cuisinées. La figure ci-contre (cf. Figure 1) représente la matrice de conception où le niveau haut (+) représente la teneur élevée de l'ingrédient tandis que le niveau bas (-) représente la teneur faible de l'ingrédient. On y retrouve la recette (Ord), la réplique de la recette (1 ou 2) et le score attribué à la recette. De plus, nous avons 8 configurations expérimentales. Ces configurations correspondent aux différentes combinaisons des niveaux haut et bas des ingrédients (farine, sel, levure) dans chaque recette, permettant ainsi d'évaluer l'effet de chaque ingrédient sur la qualité gustative de la pizza.

repl	flour	salt	bakPow	Score	Ord
1	-	-	-	5.33	7
1	+	-	-	6.99	1
1	-	+	-	4.23	4
1	+	+	-	6.61	8
1	-	-	+	2.26	5
1	+	-	+	5.75	2
1	-	+	+	3.26	6
1	+	+	+	6.24	3
2	-	-	-	5.70	7
2	+	-	-	7.71	1
2	-	+	-	5.13	4
2	+	+	-	6.76	8

Figure 1 – Matrice de conception

Après avoir dressé cette matrice, j'ai pu mettre en place mon modèle de régression de type ANOVA en prenant en compte les interactions entre les facteurs. D'après ce modèle et l'étude des coefficients, il a été possible d'identifier les facteurs significatifs, en l'occurrence la farine et la levure. Un modèle simplifié, sélectionnant les facteurs significatifs (excluant le facteur levure et les interactions), a ensuite été développé, mettant en évidence l'influence de ces deux facteurs sur le score gustatif. J'ai pu, à l'aide de mon modèle simplifié, prédire des scores pour chacune de mes conditions expérimentales (les 8 configurations possibles).

Ainsi, mes prédictions ont mis en évidence qu'une teneur élevée en farine influence positivement le score attribué à la pâte et, à l'inverse, une teneur élevée en levure influence ce score négativement. Le graphique ci-contre (cf. Figure 2) permet de visualiser ces impacts. En effet, l'on observe que si la teneur en farine (en bleu) passe du niveau bas au niveau haut, alors le score obtenu augmente de 1,23 point. À l'inverse, si la teneur en levure (en rouge) passe du niveau bas au niveau haut, alors le score diminue de 1,87 point.

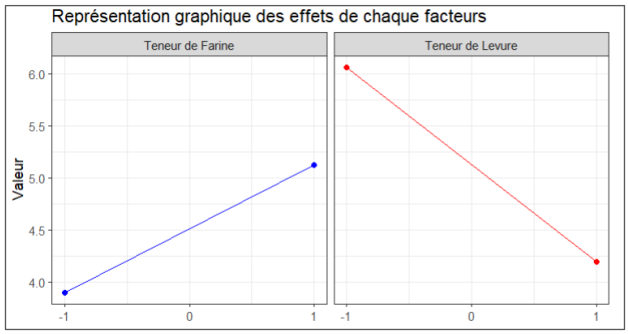


Figure 2 - Représentation des effets de la farine et de la levure

C. Conclusion

Ce projet m'a permis d'étudier des modèles statistiques plus complexes en prenant en compte simultanément plusieurs facteurs, et en examinant les effets de ces facteurs ainsi que leurs interactions. La mise en place de ce plan d'expérience a permis de réduire le nombre d'essais nécessaires. En effet, sans plan d'expériences, l'exploration des différentes combinaisons d'ingrédients aurait été laborieuse, nécessitant davantage de temps, d'argent et de matériaux.

A. Présentation

Ce projet avait pour objectif de nous confronter aux problématiques liées à l'intégration de données hétérogènes. Lors de ce travail, il a été nécessaire d'intégrer les données de 37 fichiers CSV dans un entrepôt de données en **MySQL**. L'ensemble des données recueillies sur le site *Airvivo* portaient sur des mesures d'indicateurs de la qualité de l'air (concentration d'azote, de monoxyde de carbone, etc.) et météorologiques (vitesse du vent, radiation, etc.) concernant la ville de Sheffield entre 2000 et 2022. Ce travail de groupe (3 personnes) a donné lieu à l'élaboration d'un cahier des charges concernant l'intégration des données dans un entrepôt, d'un rapport d'analyse sur la qualité de l'air de la ville ainsi que d'une soutenance orale permettant de restituer nos résultats.

B. Démarche et résolution du problème

En vue de réaliser une analyse, nous avons suivi les différentes étapes énoncées ci-dessous : conception de l'entrepôt de données, traitement et chargement des données et enfin, l'analyse.

Conception de l'entrepôt de données : Dans cette première partie du projet, nous avons réfléchi à la structure de notre base de données et identifié les ajustements requis pour la transformation de nos données. Pour cela, nous avons mis en œuvre un cahier des charges. Nous y avons défini la structure de l'entrepôt en présentant : un schéma conceptuel et relationnel, la structure des différentes tables, la définition des métadonnées, les contraintes (clés primaires et étrangères, conventions de codage, etc.).

Traitement et chargement : Pour intégrer les données dans notre entrepôt, nous avons automatisé la lecture et le traitement (transformation des données, normalisation des données, insertion au sein de tables, etc.) de ces fichiers au moyen du langage de programmation **Python** et de la bibliothèque **Pandas**. Pour cela, nous avons mis en place un notebook Jupyter documentant le code. Les données ont finalement été insérées dans notre base de données **MySQL** à l'aide de la librairie **Mysql.connector** de **Python**.

Analyse : Pour faciliter nos analyses sous **Python**, nous avons établi des vues SQL permettant de regrouper et d'agréger des données de différentes tables au sein d'une même vue. Par exemple, nous avons créé une vue « Indicateur » qui regroupe les mesures de concentration de plusieurs molécules. Cette vue nous a ensuite permis de déterminer la valeur maximale mesurée chaque jour pour chacune des molécules (cf. Figure 3). Nous avons par la suite réalisé des représentations graphiques à l'aide des librairies **Matplotlib**, **Numpy** et **SciPy** de **Python**.

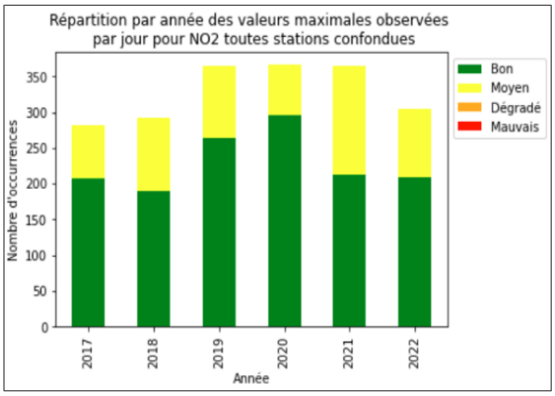


Figure 3 - Extrait d'une représentation graphique

C. Conclusion

Ce projet nous a permis de concevoir notre propre base de données et de découvrir l'existence des vues SQL. Le travail de réflexion et la mise en place du cahier des charges ont été cruciaux, permettant de clarifier nos besoins et attentes. Nous avons également soulevé une problématique essentielle dans la mise en œuvre de ce travail : l'importance des tests. En effet, après avoir constaté l'arrondi non désiré de certaines valeurs au sein de la base de données, nous avons été obligés de procéder à une seconde étape d'insertion des données. Des tests préalables sur quelques lignes auraient évité de perdre un temps considérable à réinsérer des millions de lignes dans la base de données.

SAÉ 3.03 – Description et prévision de données temporelles

A. Présentation

Lors de ce projet, nous avons examiné une série chronologique analysant le nombre mensuel de nuitées enregistrées dans les hôtels de Normandie sur la période de janvier 2011 à septembre 2023. Notre objectif était de créer un modèle prédictif pour l'année 2024 du nombre de nuitées attendu dans les hôtels de Normandie. Pour cela, nous avons travaillé en groupe de 3 étudiants sur les données de l'INSEE à l'aide du langage de programmation statistique **R**. Ce travail a abouti à la rédaction d'un rapport d'analyse suivi d'une présentation orale.

B. Démarche et résolution du problème

Le chronogramme ci-contre permet de visualiser l'entièreté de la série chronologique avec, en rouge, la prévision réalisée correspondant au nombre de nuitées attendues pour l'année 2024. On retrouve sur l'axe des abscisses les mois de chaque année depuis 2011 et, sur l'axe des ordonnées le nombre de nuitées. Pour parvenir à un tel résultat, nous avons suivi les différentes étapes présentées ci-dessous.

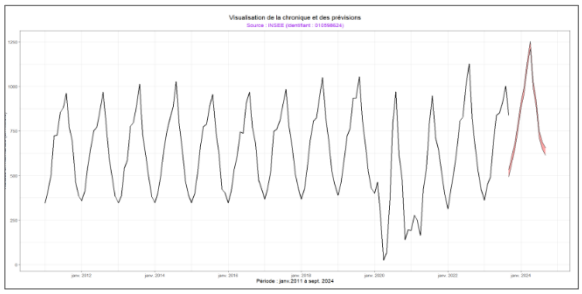


Figure 4 - Chronogramme de la série temporelle

Analyse exploratoire : Dans l'objectif de mieux comprendre nos données, nous avons visualisé la série en utilisant des visualisations telles qu'un chronogramme pour détecter les tendances et les variations saisonnières. La tendance représente l'évolution générale des observations sur une période donnée, tandis que les saisons correspondent à des schémas répétitifs. Ainsi, si l'on se réfère à la Figure 4, on remarque une tendance croissante du nombre de nuitées à partir de 2020 et la présence d'une saisonnalité d'une période d'un an (schéma répétitif) où le nombre de réservations atteint un pic chaque mois d'août.

Choix du modèle : Aux vues de la visualisation de la série temporelle, nous avons choisi un schéma additif. Dans un schéma additif, la série chronologique est modélisée comme la somme des composantes de la tendance, de la saisonnalité et d'erreurs (fluctuations aléatoires).

Création des modèles : Nous avons procédé à la décomposition de la série en isolant la tendance d'un côté et les coefficients saisonniers de l'autre. Pour ce faire, nous avons utilisé un modèle de régression linéaire pour nos deux composantes. Pour calculer la tendance, nous avons mis en place un modèle de régression linéaire par morceaux en utilisant la méthode des moindres carrés ordinaires. En ce qui concerne les saisons, nous avons soustrait la tendance estimée de notre modèle à la série temporelle initiale. En utilisant des variables indicatrices, représentant les différents moments saisonniers, nous avons pu obtenir les coefficients saisonniers. Ainsi, on observe sur la figure ci-contre une augmentation du nombre de nuitées au mois d'août par rapport à la moyenne (378 000 réservations en moyenne sur ce mois).

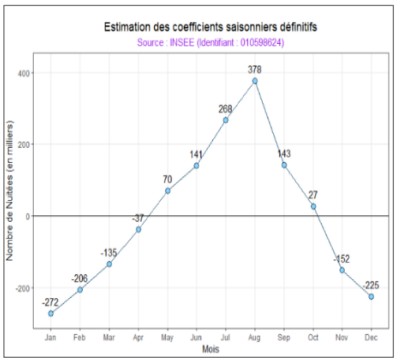


Figure 5 - Coefficients saisonniers

Vérification des modèles : Pour s'assurer de la validité du modèle, nous avons examiné les résidus pour repérer toute valeur mal ajustée et de détecter la présence d'informations non prises en compte par le modèle.

C. Conclusion

Ce projet m'a permis de prendre en main des méthodes statistiques plus complexes (décomposition de séries chronologiques) afin de réaliser des prédictions. De plus, lors de ce travail, j'ai dépassé la simple utilisation des méthodes statistiques en faisant un véritable effort de contextualisation des résultats.

A. Présentation

Ce module s'est présenté sous la forme d'enseignements et a fait l'objet d'une restitution de connaissances lors d'une séance d'examen sur feuille. Il avait pour objectif de nous sensibiliser aux enjeux de la réglementation générale sur la protection des données et aux problématiques de mise en conformité des données. Pour cela, nous avons échangé avec le Délégué à la Protection des données du Crédit Agricole de Normandie, venu spécialement pour l'occasion, afin d'échanger autour des défis et des bonnes pratiques en matière de protection des données personnelles.

B. Contenu de l'enseignement

Ces échanges avec le DPO du Crédit Agricole de Normandie ont eu deux objectifs : présenter la réglementation et partager les bonnes pratiques en matière de gestion des données.

La réglementation liée au traitement des données :

J'ai pu approfondir ma compréhension sur les missions de la Commission Nationale de l'Informatique et des Libertés (**CNIL**). En effet, la CNIL joue un rôle central dans la régulation et l'accompagnement du respect des données personnelles. Nous avons également introduit le Règlement Général sur la Protection des Données (**RGPD**), une législation européenne visant à encadrer le traitement des données au sein de l'Union européenne. Nous avons défini ce qu'est une donnée sensible, soulignant que, sauf exception, de telles données ne peuvent être traitées et analysées. Parallèlement, nous avons examiné les diverses bases légales sur lesquelles repose le traitement légitime de ces informations ainsi que l'ensemble des droits des individus, notamment le droit d'accès, à l'effacement ou encore à la limitation du traitement. En outre, nous avons évoqué les enjeux autour des transferts de données à l'étranger (par exemple, le **Data Privacy Framework**).

Les bonnes pratiques :

Nous avons également échangé autour des bonnes pratiques à mettre en place. Nous avons essayé de répondre à diverses questions : Comment traiter des données de façon responsable ? Quelles mesures de protection pouvons-nous mettre en œuvre ?

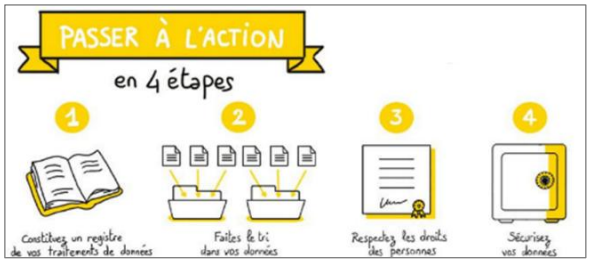


Figure 6 - Image réalisée par la CNIL montrant les 4 bonnes pratiques à suivre lors d'un projet

Pour garantir une gestion responsable des données, il existe différents points de vigilance : la durée de conservation des données, la sécurisation des informations, la mise en place de processus de tri dans les données ou encore la minimisation des données. Afin de respecter l'ensemble des réglementations durant un projet, j'ai retenu qu'il est nécessaire de débuter tout traitement par une phase de réflexion sur la licéité et la légitimité de la démarche (stratégie de **Privacy by Design**). En effet, l'analyse de données ne doit pas permettre d'identifier une personne spécifique, c'est pourquoi il est important de s'y pencher en début de projet au risque d'effectuer des traitements non conformes et donc non exploitables. À cette fin, il est conseillé de mettre en place un registre recensant l'ensemble des traitements exploitant des données (cf. Figure 6).

C. Conclusion

Ces échanges ont contribué à l'enrichissement de mon vocabulaire spécifique, un atout qui facilitera mon intégration dans le monde professionnel. De plus, cela a suscité une sensibilisation approfondie aux enjeux liés à la protection des données personnelles. Je suis désormais informé sur les ressources disponibles pour obtenir des conseils et des orientations en matière de protection des données.

SAÉ 4.01 – Expliquer ou prédire une variable quantitative à partir de plusieurs facteurs

A. Présentation

Au cours de ce projet, nous avons travaillé en groupes de trois étudiants à l'aide de **RStudio** sur l'analyse de données concernant des manchots. Les données provenaient du package « **palmerpenguins** » de **R**, qui recense des observations collectées par le Dr. Kristen Gorman sur l'archipel Palmer. L'objectif de ce projet était de créer des modèles de régression linéaire afin d'expliquer le poids d'un manchot à partir de différentes variables quantitatives et qualitatives. Nous avons développé différents modèles de diverses complexités, allant des modèles de régression linéaire simples aux modèles de régression linéaire de type ANOVA et ANCOVA.

B. Démarche et résolution du problème

Régression linéaire simple : Nous avons débuté par une analyse de la longueur de la nageoire, de la profondeur et de la longueur du bec, en évaluant leur association linéaire avec le poids des manchots. Après avoir appliqué des tests, notamment le test de Fisher, et comparé nos coefficients de détermination, nous avons conclu que la longueur de la nageoire était la variable la plus pertinente pour exprimer cette relation.

Analyse ANOVA : Nous avons ensuite procédé à une analyse de variance (ANOVA) pour étudier l'effet du sexe et de l'espèce sur le poids des manchots. Cette analyse nous a permis de déterminer que l'espèce était la variable ayant le plus d'influence sur le poids. Pour cette comparaison, nous avons utilisé la méthode du coefficient de détermination ajustée car la complexité de nos deux modèles était différente.

Analyse ANCOVA : Nous avons ensuite utilisé une analyse de covariance (ANCOVA) en intégrant la longueur de la nageoire, qui avait été sélectionnée précédemment, couplée à l'espèce ou au sexe. Cette analyse a évalué si l'inclusion de la longueur des nageoires, en association avec l'espèce ou le sexe, améliorait la capacité d'explication du poids des manchots. Nous avons utilisé des tests de Fisher pour comparer la significativité de nos modèles, tout en examinant les interactions significatives entre la longueur des nageoires et ces variables. La figure ci-contre met en évidence une interaction entre la longueur de la nageoire et l'espèce ; en effet, les droites de régression lissées ne sont pas parallèles.

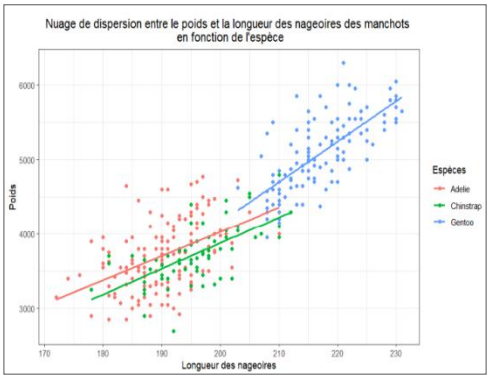


Figure 7 - Distribution entre le poids et la longueur des nageoires

Modélisation ascendante et descendante : Nous avons procédé à une sélection de modèles en utilisant une approche ascendante et descendante. Cette démarche a consisté à ajouter ou retirer sélectivement des variables explicatives du modèle de régression, en fonction de leur contribution à l'explication de la variation du poids des manchots. Cette méthode nous a permis de parvenir à un modèle final optimal, offrant un équilibre entre performance prédictive et complexité du modèle.

C. Conclusion

En résumé, ce projet m'a offert l'opportunité de maîtriser une gamme étendue d'outils statistiques, allant des différentes méthodes de comparaison de modèles (test de Student, test de Fisher, méthode AIC, etc.) aux techniques d'évaluation de la validité du modèle (normalité, homogénéité, gestion des valeurs atypiques). Travailler sur des modèles de diverses complexités m'a permis de mieux appréhender l'utilisation de ces méthodes.

A. Présentation

La cohorte **AGRICAN** est une cohorte prospective portant sur une population d'affiliés au régime agricole d'assurance maladie. Elle s'intéresse à la santé des agriculteurs et des secteurs connexes à l'agriculture. Pour ce projet, nous avons travaillé sur une sous-cohorte (près de 60 000 agriculteurs) de ce panel. Ces données portaient sur les activités agricoles pratiquées par les agriculteurs telles que l'élevage de bovins, de porcs, la culture de maïs, etc. L'objectif était de reconstruire la carrière des agriculteurs dans le but de regrouper les parcours agricoles similaires au moyen d'une méthode statistique de classification (méthode de **clustering des K-means**). Ce travail s'est effectué par groupe de deux étudiants sur le langage de programmation **R** et a donné lieu à un rapport d'analyse édité au moyen de R Markdown.

B. Démarche et résolution du problème

Notre première étape a été de prendre connaissance du questionnaire à partir duquel les réponses ont été recueillies. Nous avons pris soin de l'annoter afin de bien comprendre l'ensemble des variables disponibles dans le jeu de données. Ensuite, nous avons mis en place un script automatisé permettant de calculer la durée totale de la carrière de chaque agriculteur et la part de chaque activité sur ce temps de carrière (ratio d'activités). Cette étape a demandé un effort considérable en termes d'automatisation, nécessitant la mise en place de différentes fonctions pour traiter un ensemble important de variables.

Nous avons ensuite réalisé une analyse en composantes principales afin de mettre place un clustering en utilisant la méthode des **K-means** pour regrouper les agriculteurs en clusters (groupes d'individus) présentant des parcours agricoles similaires. L'algorithme des K-means itère successivement en ciblant des points de données comme centroïdes et en calculant leurs distances par rapport à tous les autres points. Ce processus se répète jusqu'à ce que la variance intra-cluster (somme des distances entre les points) ne diminue plus de manière significative à chaque itération, indiquant ainsi que les centroïdes sont stabilisés et que le nombre optimal de clusters est alors atteint. Enfin, nous avons examiné les caractéristiques communes des agriculteurs dans chaque cluster.

Le tableau (cf. Figure 8) ci-contre est un extrait des différents clusters obtenus. Dans la colonne « Moyenne », nous trouvons les ratios d'activité. Les ratios significativement inférieurs à la moyenne sont représentés en rouge, et ceux significativement supérieurs sont représentés en vert. Par exemple, pour l'élevage de bovins, la moyenne observée à travers tous les clusters était de 61,40 %. Cela signifie que, en moyenne, les agriculteurs de la cohorte consacrent 61,40 % de leur temps de carrière totale à l'élevage de bovins, que ce soit comme activité principale ou en conjonction avec d'autres pratiques agricoles durant cette période.

Table 4: Caractéristiques des clusters

	Moy c8	v-test8	Moy c6	v-test6	Moyenne
Bovins	9.25	-55.55	89.59	21.53	61.40
Moutons/chèvres	0.58	-11.53	5.74	-0.08	5.79
Cochons	1.26	-18.23	65.15	58.39	12.95
Chevaux	3.24	-13.00	61.13	59.00	11.14
Volailles	3.11	-15.78	58.38	45.74	14.07
Vigne	95.63	71.46	34.68	4.56	28.73
Maïs	4.5	-31.84	25.92	-5.17	32.20
Bettraves	1.12	-19.55	63.71	49.81	14.87
Tournesol	0.55	-10.94	0.89	-7.15	4.45
Colza	0.34	-12.54	7.08	2.31	5.70
Tabac	0.78	-10.23	5.99	1.62	5.05
Arboriculture	2.16	-13.06	24.58	16.59	10.25
Prairies	14.39	-44.72	87.66	22.11	57.75
Autres.légumières	0.66	-11.50	10.62	6.28	6.31
Blé.ou.orge	7.27	-41.32	83.92	26.61	47.66
Cultures.sous.serres	0.27	-8.00	0.81	-4.60	3.01
Pois.fourragers	0.09	-9.84	1.36	-4.40	3.47
Pommes.de.terre	2.61	-17.77	70.59	52.63	15.86

C. Conclusion

En conclusion, ce projet a permis de reconstruire les carrières des agriculteurs et de les rassembler en groupes homogènes selon des parcours similaires. Il nous a fallu traiter une grande quantité de données, nécessitant une approche méthodique pour effectuer nos ratios. De plus, l'utilisation de la méthode de clustering des **K-means** a été une expérience enrichissante, me permettant de clarifier les mécanismes sous-jacents du clustering en passant par les étapes de sélection des composantes, du choix du nombre de clusters optimal et d'analyse.

A. Présentation

Ce projet a été réalisé en groupe de trois étudiants, sous la supervision d'un doctorant tuteur du centre François Baclesse (Centre de lutte contre le cancer à Caen). L'objectif était de produire un article scientifique sur l'étude du cancer du poumon au sein de la cohorte **AGRICAN** (AgriCancer) qui recense 180 000 affiliés à la MSA française. En fin de projet, nous avons présenté nos résultats lors d'une soutenance orale devant un jury composé des encadrants du projet.

B. Démarche et résolution du problème

Pour réaliser ce projet nous avons commencé par effectuer des recherches scientifiques sur le cancer du poumon, ainsi que sur l'incidence du cancer chez les agriculteurs en France et dans le monde, afin de contextualiser notre étude. Par la suite, nous avons eu accès à la base de données anonymisée de la cohorte AGRICAN fournie par le centre François Baclesse. Nous avons préparé ces données sous SAS en filtrant les observations pertinentes et en sélectionnant les variables qui nous intéressaient (Figure 1).

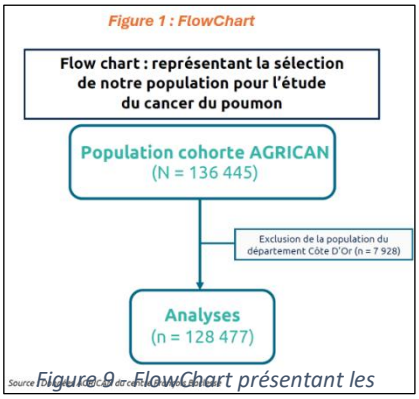


Figure 9 : FlowChart présentant les individus analysés

Par la suite, nous avons réalisé une analyse exploratoire sur nos deux groupes d'étude (malades et non malades) avec le **langage R** en utilisant un test du Khi-deux pour les variables qualitatives et un test de Student pour les variables quantitatives. Nous avons décrit les caractéristiques intrinsèques de notre échantillon, notamment la répartition des cas de malades et de non-malades selon l'âge, le sexe, la consommation de tabac ou encore les activités agricoles pratiquées.

Enfin, nous avons modélisé nos données à l'aide d'un **modèle de Cox** réalisé sous le logiciel **SAS** afin d'évaluer l'influence des activités agricoles et des facteurs socio-démographiques en fonction du temps de suivi des individus dans la cohorte. Ce modèle nous a permis de standardiser l'analyse en tenant compte de l'âge, un facteur de confusion susceptible de biaiser les résultats. L'interprétation des résultats a permis d'identifier des facteurs protecteurs ainsi que des facteurs de risque. L'ensemble de ces analyses a été synthétisé sous la forme d'un article scientifique.

C. Conclusion

Finalement, ce projet nous a permis de nous initier à la recherche scientifique tout en nous sensibilisant aux risques liés au cancer du poumon. Ce cancer est principalement influencé par l'environnement de l'individu, les principaux facteurs de risque étant le tabagisme, l'exposition au radon (un gaz radioactif naturel issu de la décomposition de l'uranium présent dans le sol) et la pollution atmosphérique.

Concernant notre étude, plusieurs axes d'amélioration sont envisageables, notamment une meilleure prise en compte des facteurs de confusion que nous avons identifiés. L'intégration de modèles plus complexes permettrait d'affiner l'analyse et d'obtenir des résultats plus robustes.

SAÉ 5.02 – Migration de données vers ou depuis un environnement NoSQL

A. Présentation

Ce projet individuel avait pour objectif de nous familiariser avec les bases de données NoSQL, et plus particulièrement les bases orientées documents. Dans le cadre de ce projet, j'ai travaillé sur l'intégration de 100 000 commandes issues d'un site d'e-commerce brésilien dans une base de données **MongoDB**. Ce travail a abouti à la création d'un notebook Jupyter contenant mon code **Python** documenté, ainsi qu'un rapport détaillant notre démarche et nos analyses.

B. Démarche et résolution du problème

Pour mener à bien ce projet, j'ai d'abord mis en place un schéma relationnel permettant d'organiser les données issues des fichiers CSV fournis. Ce schéma comprend 7 collections correspondant aux entités clés du dataset : commandes, paiements, clients, vendeurs, produits, articles commandés et avis clients (Figure 2).

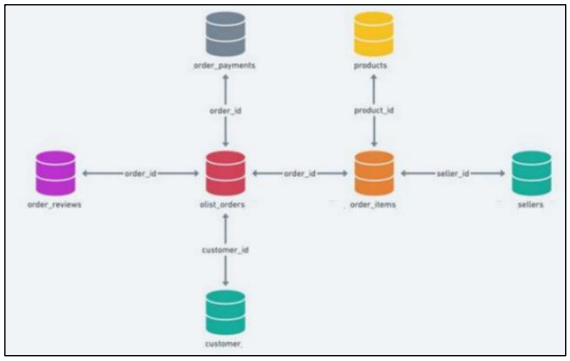


Figure 10 - Schéma relationnel de la base de données

Dans un premier temps, j'ai utilisé la bibliothèque **PyMongo** disponible en Python pour me connecter à MongoDB. Ensuite, j'ai créé la structure des 7 collections en définissant les clefs de jointure et en les indexant afin d'optimiser les performances lors du requêtage. Une fois la structure en place, j'ai procédé à l'importation des 7 fichiers CSV dans MongoDB. Chaque fichier a été chargé dans sa collection respective.

Avec la base de données constituée, j'ai effectué plusieurs requêtes impliquant une ou plusieurs collections afin d'extraire des informations répondant à diverses problématiques, telles que l'identification des vendeurs les mieux notés, des États brésiliens les plus prolifiques en termes de commandes, ou encore des délais de livraison. Enfin, j'ai présenté mes analyses à l'aide de la bibliothèque **Matplotlib** du langage Python, en réalisant des graphiques pour présenter les résultats au sein d'un rapport.

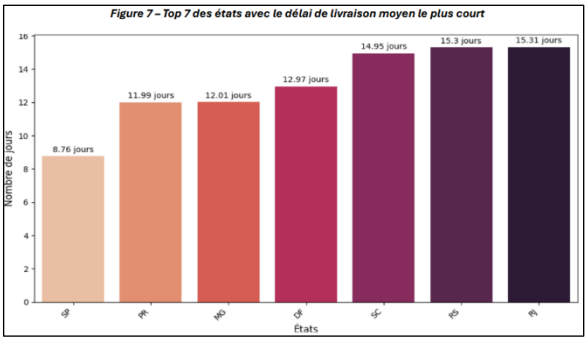


Figure 11 – Représentation graphique des états avec les délais de livraison les plus courts

C. Conclusion

Finalement, ce projet m'a permis de mieux comprendre le fonctionnement de MongoDB et d'explorer plusieurs fonctionnalités essentielles, notamment les opérations de requêtage avec la fonction **find()** et les jointures avec l'opération d'agrégation **$lookup**. J'ai également pris conscience de l'importance du schéma relationnel, même dans une base NoSQL, et de la nécessité d'une bonne structuration des données. L'un des principaux défis a été la gestion des jointures entre les collections, qui étaient nombreuses et parfois complexes. De plus, au début du projet, je n'avais pas indexé mes données, ce qui ralentissait considérablement les requêtes. Après plusieurs recherches, j'ai découvert le principe d'indexation et son impact sur les performances, ce qui m'a permis d'optimiser significativement mes requêtes.

A. Présentation

Lors de ce projet individuel j'ai travaillé sur près de 100 000 avis clients collectés sur le site d'e-commerce brésilien : Olist. L'objectif était de développer **deux modèles de clustering** : l'un basé sur la méthode **Bag of Words** et l'autre sur l'approche **TF-IDF**. Pour chacun des modèles, les avis ont été classés en quatre catégories : très satisfait, satisfait, insatisfait et mécontent. Le projet s'est conclu par la rédaction d'un Jupyter Notebook documentant l'ensemble du code **Python**, ainsi qu'une analyse détaillée dans un rapport.

B. Démarche et résolution du problème

La première étape du projet a consisté à nettoyer les avis clients en supprimant les valeurs manquantes, les doublons, les caractères spéciaux, la ponctuation, les emojis et les stop words en portugais à l'aide de **nltk** et **spaCy**. Les commentaires ont été convertis en minuscules, puis un échantillon de **30 000 avis** a été extrait pour l'analyse.

Afin de pouvoir classifier les avis clients à l'aide d'un réseau de neurones, il était nécessaire de convertir les données textuelles en une représentation numérique compréhensible par le modèle. Pour cela, deux méthodes ont été comparées : Bag of Words (BoW) et TF-IDF.

- **Bag of Words (BoW)** repose sur un comptage de la fréquence des mots dans chaque avis, sans prendre en compte leur importance ou leur contexte.
- **TF-IDF (Term Frequency - Inverse Document Frequency)**, attribue un poids aux mots en fonction de leur fréquence dans un avis, tout en réduisant l'importance des termes trop courants dans l'ensemble du corpus. Cela permet de mieux distinguer les mots les plus significatifs.

Ces deux méthodes produisent des matrices de grande dimension, pour réduire la taille de ces matrices, la technique de Truncated SVD (Singular Value Decomposition) a été appliquée. Les vecteurs réduits obtenus après Truncated SVD ont ensuite été utilisés comme entrées pour le réseau de neurones. Afin d'entraîner et d'évaluer ce modèle, les données ont été divisées en deux échantillons : 80 % pour l'échantillon d'apprentissage et 20 % pour l'échantillon de test.

Une fois les modèles entraînés, j'ai procédé à une évaluation approfondie de leurs performances. J'ai ainsi calculé l'ARI (Adjusted Rand Index), analysé la matrice de confusion pour comparer les performances des deux techniques de vectorisation (BoW et TF-IDF) et tracé les courbes ROC afin d'évaluer la capacité du modèle à discriminer correctement les classes (Figure 4).

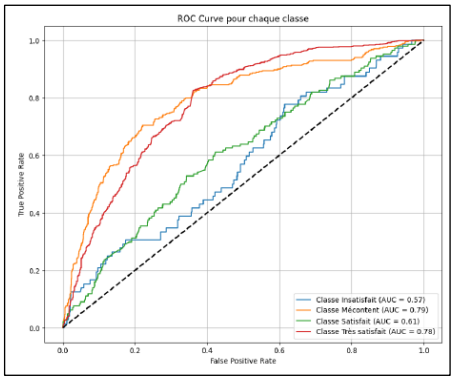


Figure 12 - Courbe ROC avec la vectorisation des données au moyen de TF-IDF

C. Conclusion

Ce projet m'a permis d'explorer les méthodes de traitement du langage naturel et de mieux comprendre leurs défis. La principale difficulté résidait dans la gestion des commentaires, certains commentaires comportaient des fautes d'orthographe, et la présence d'abréviations constituait également un défi. De plus, l'analyse des résultats a montré que le réseau de neurones avait du mal à distinguer les classes "satisfait" et "insatisfait", tandis qu'il identifiait plus précisément les avis "très satisfait" et "mécontent". Une approche réduisant le clustering à deux classes au lieu de quatre aurait peut-être permis d'améliorer la performance du modèle.

SAÉ 6.01 – Modélisation statistique pour les données complexes et le Big Data

A. Présentation

Ce projet s'est déroulé par groupes de 3 étudiants. Il a visé à analyser l'influence de la carrière professionnelle des agriculteurs sur l'apparition d'une maladie au sein de la cohorte AGRICAN (AgriCancer) qui recense 180 000 agriculteurs affiliés à la MSA française. Pour réaliser cette étude, nous avions à notre disposition 10 413 agriculteurs et 18 variables concernant des activités d'élevage ou de culture. Ce projet s'est terminé par le rendu d'un rapport et de notre script réalisé au moyen du **langage R**.

B. Démarche et résolution du problème

Dans un premier temps, nous avons réalisé une analyse exploratoire des données afin d'identifier d'éventuelles valeurs aberrantes et de mieux comprendre la structure du jeu de données. Pour identifier les variables les plus significatives en vue de la modélisation, nous avons appliqué deux méthodes de sélection sur 50 échantillons Bootstrap :

- **Méthode 1** : **Régression logistique**. Nous avons réalisé une régression logistique multiple sur les 18 variables explicatives pour chaque échantillon Bootstrap. Les variables significatives ont été sélectionnées sur la base des p-values obtenues. Nous avons retenu les 5 variables les plus significatives.
- **Méthode 2** : **Méthode de Backward**. Cette méthode consiste à éliminer progressivement les variables les moins significatives. Finalement, nous avons retrouvé les 5 mêmes variables que précédemment.

Ainsi, nous avons modélisé les données à partir de deux bases de données. La base de données réelle en ne retenant que les 5 variables significatives et une base de données réduite (via une ACP). Par la suite, nous avons réalisé 40 échantillons d'apprentissage (80% des données) avec la méthode Bootstrap et 40 échantillons de test (20% des données) sur les deux bases de données. Pour chacun des modèles, nous avons comparé les résultats obtenus avec la base réduite et la base de données réelle.

- **Régression logistique :** Une régression classique, une régression avec la méthode bagging et une régression avec la méthode boosting.
- **Arbres de décision :** classique, boosting (Adaboost M1), forêts aléatoires.
- **Réseaux de neurones :** Nous avons testé trois à quatre architectures différentes pour optimiser la performance du modèle, en modifiant le nombre de couches intermédiaires et le nombre de neurones dans ces couches.

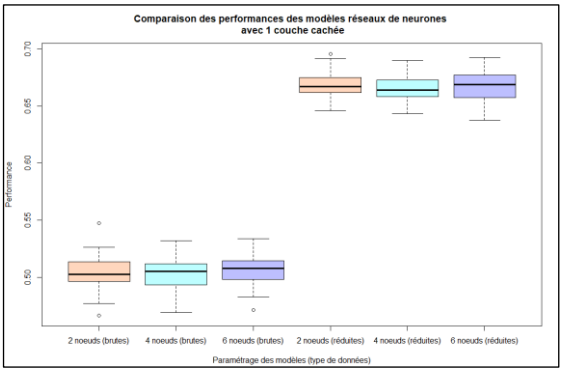


Figure 13 - Comparaison des performances des réseaux de neurones avec les données réelles versus les données réduites

C. Conclusion

Finalement, parmi les différentes méthodes testées, la régression logistique et les arbres de décision ont offert de bonnes performances sur la base des données réduites avec un temps de calcul raisonnable. En revanche, l'utilisation des réseaux de neurones s'est révélée plus complexe. L'ajustement des paramètres, notamment le paramètre du stepmax (nombre maximal d'itérations pour la convergence), a nécessité plusieurs essais afin d'éviter les problèmes de non-convergence ou de surajustement.